Evaluation of in situ sea surface temperatures for use in the calibration and validation of satellite retrievals

Feng Xu[1,2] and Alexander Ignatov[1]

[1] In situ sea surface temperatures (SSTs) are used for calibration and validation of satellite retrievals. This study analyzes three in situ data sets from the National Centers for Environmental Prediction (NCEP) Global Telecommunication System (GTS), the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) release 2.4, and the U.S. Global Ocean Data Assimilation Experiment/Fleet Numerical Meteorology and Oceanography Center (FNMOC). Comparisons show that most reports in the ICOADS and FNMOC are of the same origin as NCEP GTS. Quality control (QC) information is either unavailable (NCEP), not well documented (FNMOC), or nonuniform (ICOADS, FNMOC). Preliminary QC was implemented in this study and uniformly applied to all data sets. All analyses are stratified by major types of in situ platforms including ships, drifters, and moored buoys, the latter being further subdivided into tropical and coastal. Ships overwhelmingly prevailed before 1990 but then declined, whereas the number of drifters significantly increased, as did their reporting density. Although both platforms sample the full SST range well, drifters cover the global ocean much more uniformly than ships. Statistical analyses are performed on the in situ SST anomalies with respect to daily Reynolds and daily Pathfinder. Different global mean biases are observed for different platform types (e.g., ~+0.03 K for drifters and tropical moorings and ~+0.15 K for ships, with respect to Reynolds SST), suggesting existence of cross-platform biases that need to be reconciled. Root mean square (RMS) errors of the four types of in situ data have been estimated via three-way analyses proposed in O'Carroll et al. (2008). The geographical distributions of RMS errors in Pathfinder, Reynolds, and in situ SSTs show distinct spatial patterns, which require further understanding and remediation.

1. Introduction

[2] Since the early 1980s, the National Environmental Satellite, Data, and Information Service (NESDIS) has routinely generated global sea surface temperature (SST) products from the Advanced Very High Resolution Radiometer (AVHRR) sensors onboard NOAA and MetOp satellites. Multichannel SST and nonlinear SST techniques are employed in retrievals [e.g., *McClain et al.*, 1985; *Walton et al.*, 1998]. The "ground truth" in situ SSTs are used to initially calibrate the satellite SST algorithms (i.e., derive coefficients of the regression equations) and then to continuously validate retrievals (i.e., monitor global statistics of "satellite minus in situ" SST differences). This process is referred in the satellite SST community as "Cal/Val."

[3] The NESDIS heritage Cal/Val uses in situ SSTs available from the National Centers for Environmental Prediction (NCEP) Global Telecommunication System (GTS) data. In situ observations include measurements from ships and drifting and moored buoys. On the basis of some earlier analyses [*Strong and McClain*, 1984], only data from drifters and tropical-moored buoys have been utilized in satellite Cal/Val, whereas ships and coastal moorings have been excluded. These Cal/Val practices are currently being revisited and redesigned at NESDIS to meet the new SST product requirements.

[4] A new AVHRR SST product became operational at NESDIS in May 2008, generated by the Advanced Clear-Sky Processor for Oceans (ACSPO) system jointly developed by the NESDIS Center for Satellite Applications and Research (STAR) and Office of Satellite Data Processing and Distribution (OSDPD). It is envisioned that in the future, ACSPO will be employed to consistently reprocess AVHRR/2 and /3 data back to NOAA-7, which was launched in 1981. The STAR/OSDPD SST team is also actively involved in the development and Cal/Val of the two next-generation global SST products, one from the Visible/Infrared Imager Radiom-

---

[1]Center for Satellite Applications and Research, National Environmental Satellite, Data and Information Service, NOAA, Camp Springs, Maryland, USA.
[2]CIRA, Colorado State University, Fort Collins, Colorado, USA.

eter Suite (VIIRS) to be flown onboard the National Polar-orbiting Operational Environmental Satellite System (NPOESS) and the other from the Advanced Baseline Imager (ABI) onboard the Geostationary Operational Environmental Satellite R-Series (GOES-R). Careful inventory and in-depth understanding of the available in situ SSTs are key to the improved Cal/Val of the historical data records as well as the new generation SST products from AVHRR, VIIRS, and ABI.

[5] The Group for High-Resolution SST (GHRSST, www.ghrsst.org) has identified that in situ SSTs play a critical role in product validation, bias correction, and derivation of the Single-Sensor Error Statistics [*Donlon et al.*, 2007]. In situ data are an essential element of the GHRSST monitoring, diagnostic, and verification systems, including the High Resolution Diagnostic Data Set (www.hrdds.net) and the GHRSST Match-up Database (www.medspiration. org/tools/mdb). Careful assessment and evaluation of in situ SSTs undertaken in this study thus directly contributes to the GHRSST project.

[6] This paper pursues four major objectives toward this goal. First, it explores two other in situ SST data sets in addition to NCEP GTS data currently employed at NESDIS: the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) release 2.4 (since 1980) and the US Global Ocean Data Assimilation Experiment (GODAE) Fleet Numerical Meteorology and Oceanography Center (FNMOC) (since 1998). The three data sets are cross-evaluated with the objective to determine their relative value for SST Cal/Val tasks. Second, the quality control (QC) information available on the ICOADS and FNMOC data is evaluated. The NCEP data arrive from the GTS stream without quality control, undergoing in-house QC at NESDIS before their use in Cal/Val. Third, it systematically compares in situ SSTs from ships, drifters, and tropical and coastal moorings. Exploring various in situ SSTs is critically important to fully utilize their potential for Cal/Val and extend satellite SST records back in time to the early 1980s, when ships were the major source of in situ data. Finally, the four types of in situ measurements are analyzed in terms of their coverage and systematic and random errors.

[7] Note that for comparison purposes in this study, the QC information available in the ICOADS and FNMOC data sets was not used. Instead, basic QCs developed at NESDIS (including checks for geolocation, duplicate records, tracks, and outliers) were implemented and consistently applied to all SSTs, which removes 10% to 15% of data.

[8] The major in situ data sets and platform types analyzed in this study are described in section 2. QC information available in ICOADS and FNMOC data sets is analyzed in section 3, followed by a description of the current NESDIS QC procedure adopted in this study. Comparisons among the NCEP, ICOADS, and FNMOC data sets, as well as among the different types of in situ sensors, are presented in section 4. Coverage and error characterization analyses of in situ SSTs are given in section 5. Section 6 concludes the paper and discusses future work.

2. Data

[9] Major sources of in situ SSTs are ships and buoys. Ship reports are mainly from the merchant ships participating in the World Meteorological Organization (WMO) Voluntary Observing Ship (VOS) program and to a lesser extent from the research vessels. There are two major classes of buoys, drifters and moored, which are deployed and operated by different countries and organizations under various national and international programs and projects, including the Global Drifter Program (www.aoml.noaa.gov/phod/dac/gdp.html), the Triangle Trans-Ocean Buoy Network (www.jamstec. go.jp/jamstec/TRITON/real_time/php/top.php), the Tropical Atmosphere Ocean (www.pmel.noaa.gov/tao/index.shtml) [*McPhaden et al.*, 1998], the Prediction and Research Moored Array in the Tropical Atlantic (www.pmel.noaa. gov/pirata/) [*Servain et al.*, 1998; *Bourles et al.*, 2008], the Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction [*McPhaden et al.*, 2009], and the National Data Buoy Center moored buoys (www.ndbc. noaa.gov), to name the major ones.

[10] In addition, there are also some SST reports from NOAA Coastal-Marine Automated Network stations (www. ndbc.noaa.gov/cman.php). However, the number of stations and their corresponding observations is very small, and they cover only a very limited geographical domain. Furthermore, they are located in coastal areas, which are usually too shallow and dynamic to obtain stable SST measurements for the use in satellite Cal/Val. Therefore, only reports from ships and buoys are analyzed in this study. The ships, drifters, and tropical- and coastal-moored buoys are analyzed separately to systematically cross-evaluate the four categories in historical perspective.

2.1. Collection of In Situ Data Via ARGOS and Distribution Via GTS

[11] In situ measurements over oceans are taken by a wide variety of sensors (for instance, drifting buoys are produced by four or five major manufacturers, (W. J. Emery, personal communication, 2009)). Many in situ platforms (particularly drifters and moorings) follow a standard protocol of data collection and transmission called Argos (www.argos-system.org). Data are first collected and stored onboard in situ platforms and then transmitted, via Argos-certified transmitters, to the NOAA (and more recently, MetOp) polar satellites at the time of their overpass. The data are then retransmitted to the ground each time the satellite passes over receiving stations. To minimize the delays, the data are also continuously broadcast to the ground and captured by the nearly 50 regional reception stations when in the satellite field of view. All downlinked data are subsequently retransmitted by the receiving stations to the two world processing centers with redundant operations. The data are processed in real time and further distributed to users via GTS, which is implemented and operated by the national meteorological services of WMO member countries and several international organizations. The GTS is a coordinated global system of telecommunication facilities and arrangements for the rapid collection, exchange, and distribution of observations and processed information within the framework of the World Weather Watch [*World Meteorological Organization*, 2007].

[12] Note that reports from some platforms are not collected via ARGOS. For example, VOS ship observations are collected via radio telephony or INMARSAT transmission (www.bom.gov.au/jcomm/vos/vos.html). For some other platforms (e.g., moored buoys) GTS may report reduced density

Table 1. NCEP, FNMOC, and ICOADS Data Sets Used in This Study

	NCEP	FNMOC	ICOADS
Availability	Jan 1991–present	Sep 1998–present	Sep 1981–Jul 2007
Web access	ftp://ftp.cdc.noaa.gov/Public/icoads/ncep_obs/	http://usgodae.fnmoc.navy.mil/pub/outgoing/fnmoc/data/ocn/sfcobs/	http://dss.ucar.edu/datasets/ds540.0/data/imma0_form.php
Accessibility	Near real time, 2 day delay.	Near real time, 2–8 days to accumulate a complete record.	Static archive (renewable at the time of new ICOADS releases).
Data sources	GTS	GTS plus other unspecified ship and buoy data.	Before Dec 2004 (ships, before Dec 1997): Delayed Mode (from various sources). Jan 2005 (ships, Jan 1999) and onward: NRT (GTS only).
Exceptions	Ship ID masked since Dec 2007.	Documentation not available in public domain.	-
Version/release information	N/A	N/A	Release 2.4
Quality control	N/A	Continuous quality indicator [0,1]. QC upgraded to v. 2 in Apr 2004.	Trimming flags plus NCDC composite QC flags.
Precision	0.1 K	0.1 K	0.1 K

data. Such missing data are later added to the ICOADS and FNMOC as they become available.

2.2. QC and Accuracy of In Situ SSTs

[13] Reportedly, root mean squared errors (RMSE) of quality-controlled in situ SSTs range from 0.3 to 0.5 K for buoys (which is comparable with accuracy of AVHRR SST) to 1.0 to 1.3 K for ships, because of large differences in sensors and data collection practices as discussed, for example, in *Strong and McClain* [1984], *Kent et al.* [1999], *Emery et al.* [2001], *Kent and Challenor* [2006], *Reynolds et al.* [2007], and *O'Carroll et al.* [2008]. The accuracies are much worse if no QC is applied, mainly because of a small fraction of large outliers. Recall that in situ SSTs are "data of opportunity" not specifically intended for use in satellite Cal/Val [*Emery et al.*, 2001].

[14] Different platforms employ a large variety of sensors, which operate in a wide range of often hostile environments and use different measurement protocols. Furthermore, buoys are left unattended for extended periods of time, and ships involve a certain degree of human-related impacts in data collection and transmission [e.g., *Strong and McClain*, 1984; *Kent et al.*, 1999; *Emery et al.*, 2001; *Kilpatrick et al.*, 2001; *Reynolds et al.*, 2002, 2007]. As a result, the quality of in situ data is nonuniform in space and time. In contrast to sensor-to-sensor noise in in situ data, the satellite infrared data are measured by only a few well-calibrated sensors but subject to large-scale artifacts, because of a nonuniform sampling in space and time and clear-sky biases, variations in satellite observation geometry and atmospheric water vapor, and residual and ambient cloud effects. Currently, in situ observations are processed and archived by NCEP, ICOADS, and FNMOC using different protocols and provided to users in different formats and with various latencies. The QC procedures employed are not fully unified and not always well documented, resulting in a limited and suboptimal use of the available quality flags and indicators for satellite Cal/Val [*Donlon et al.*, 2007].

2.3. In Situ SST Data Sets Used in This Study

[15] Three in situ SST data sets are studied here: the NCEP GTS real time, FNMOC surface observation, and ICOADS release 2.4. Table 1 summarizes the information about these data sets.

[16] Archived NCEP data are available dating back to January 1991. They originate exclusively from GTS and are processed by NCEP in near real time. No QC or pre-screening is routinely applied. Data have been grouped into monthly files, and the file for the most recent month in the archive is updated daily with a 2 day lag. Records include SST and some other weather data, as well as location, time, and platform ID. Platform type is "ship," "drifter," "moored buoy," or "Coastal-Marine Automated Network (C-MAN) station."

[17] FNMOC data are being processed and updated in near real time but require 2 to 8 days to acquire a complete record. The bulk of data is from GTS, plus a few extra unspecified sources of data. The data are available from September 1998 onward. A continuous quality indicator (QI) is provided.

[18] ICOADS release 2.4 archives ocean observations from 1784 to July 2007. Only data during the satellite era from 1 September 1981 onward are analyzed in this study. ICOADS is stratified into two archives: the delayed mode (DM) before December 2004 (for ships, before December 1997) and the real time (RT) from January 2005 (for ships, from January 1998) onward [*Worley et al.*, 2005; *Woodruff et al.*, 2008]. The RT archive is exclusively from the NCEP GTS data, but in a more complete form (based on the Binary Universal Form for the Representation of meteorological data (BUFR) format) (S. D. Woodruff, personal communication, 2009). Additionally The DM archive includes non-GTS reports from a variety of sources provided by different countries and organizations. A set of quality flags (QF) are provided [*Worley et al.*, 2005; *Woodruff et al.*, 1998; *Slutz et al.*, 1985]. Information retained from the original reports is also available, which may be used to trace the processing history back to the origin. As of this writing, a new ICOADS release 2.5 became available, which, among other enhancements, extended the DM processing to present time and applied a more consistent QC to all data in the satellite era (S. D. Woodruff, personal communication, 2009) [*Worley et al.*, 2009].

2.4. Number of Platforms, Observations, and Coverage

[19] This section summarizes some basic statistics of the three data sets stratified by platform types. All analyses have been carried out on a monthly basis in order to provide fine
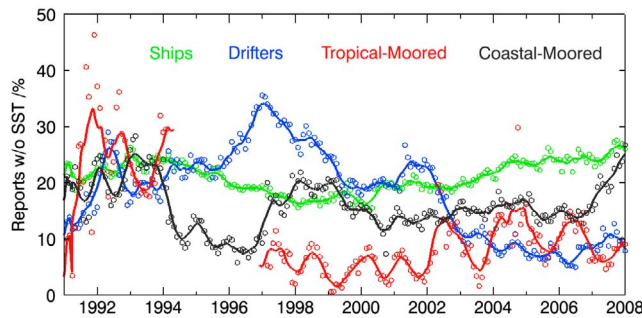
Figure 1. Percentage of reports without SST data in monthly NCEP GTS files. Fraction of ships not reporting SST is ~20%–25%. For drifters in recent years, the fraction is ~10% (but it was as high as 20%–30% prior to 2002). Fraction of moored buoys is lower for tropical moorings and changes seasonally. Note that missing tropical-moored reports in the NCEP GTS data from 23 February 1994 to 1 March 1997 are due to misclassification of tropical buoys as drifters.

resolution for the time series, while preserving a sufficient sample size for statistics. For analyses of basic statistics in this section, all numbers in the data sets were counted "as is." In particular, no attempt was made to exclude possibly erroneous records in NCEP (while ICOADS and FNMOC may have screened out some reports by duplicate removal and track checks, as discussed below in section 3).

[20] Note that the ICOADS and FNMOC data were downloaded via a Web interface with an SST subsetting function, and, therefore, all downloaded reports include SSTs, whereas the NCEP files contain all GTS reports, including those without SST. Figure 1 shows time series of the fraction of NCEP reports with missing SST. On average, 10% to 30% of in situ data do not report SST. In what follows, only platforms and reports with valid SST will be shown and discussed when using NCEP GTS data.

[21] Figure 2 shows time series of the number of platforms, based on the platform identity, or ID, provided in each data set. Only platforms with two or more reports per month are shown in Figure 2 (left), and those with one single report per month are shown in Figure 2 (right). The "single reporters" are mostly ships, which numbered anywhere from 1000 to 2000 before 2002, likely due to unreliable ship IDs. The number of single reporters is similar between NCEP and ICOADS. The drop-off in mid-2002 suggests that the treatment of the single reporters has improved in the GTS data. In FNMOC, the treatment of single reporters did not improve until early 2006. Including rarely reporting platforms in the count may artificially inflate the number of reporting platforms. On the other hand, the numbers shown in Figure 2 (left) may be underestimated because generic IDs were used to represent a group of ships or anonymous platforms, such as ship and buoy.

[22] Concentrating now on the Figure 2 (left), there were ~5000 SST-reporting ships per month before the 1990s, while buoys were only present in single digits. Following the onset of the Tropical Ocean Global Atmosphere program in 1985, buoys have been deployed increasingly [*McPhaden et al.*, 1998], while fewer and fewer ships contribute to the observations [cf. *Kent et al.*, 2006]. Most recently, the number of ships and drifters reporting SST has been ~1,500 and ~1,300 per month, respectively, while the corresponding number of moored buoys is ~300, of which ~100 are in the tropics and ~200 in coastal areas. Note that since December 2007, the number of ships in the NCEP data set is unknown because ship ID is no longer available because of the security concerns by ship owners (D. Stokes, personal communication, 2008). The apparent drop-off in the ICOADS ship coverage record at the end of 1997 is due to a change in the source of ship data in ICOADS from the DM to RT archive [see *Reynolds et al.*, 2002; http://icoads.noaa.gov/news_fig1.html].

[23] Corresponding numbers of SST reports are plotted in Figure 3 (left). Ship reports have steadily declined and, as of
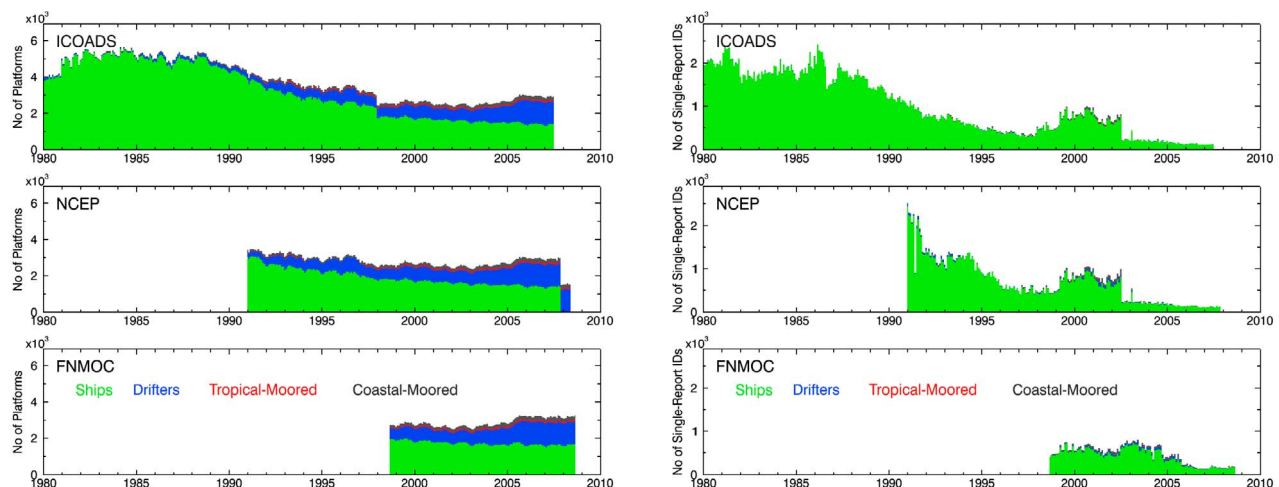

Figure 2. (left) Monthly number of SST-reporting platforms in the three data sets (based on unique IDs). Only number of platforms with monthly number of reports >1 are shown. (right) Number of platforms with only one report per month. Note that NCEP stopped reporting ship IDs beginning in December 2007. Also note that the Figure 2 (left) and Figure 2 (right) scales for the *y* axis are different.
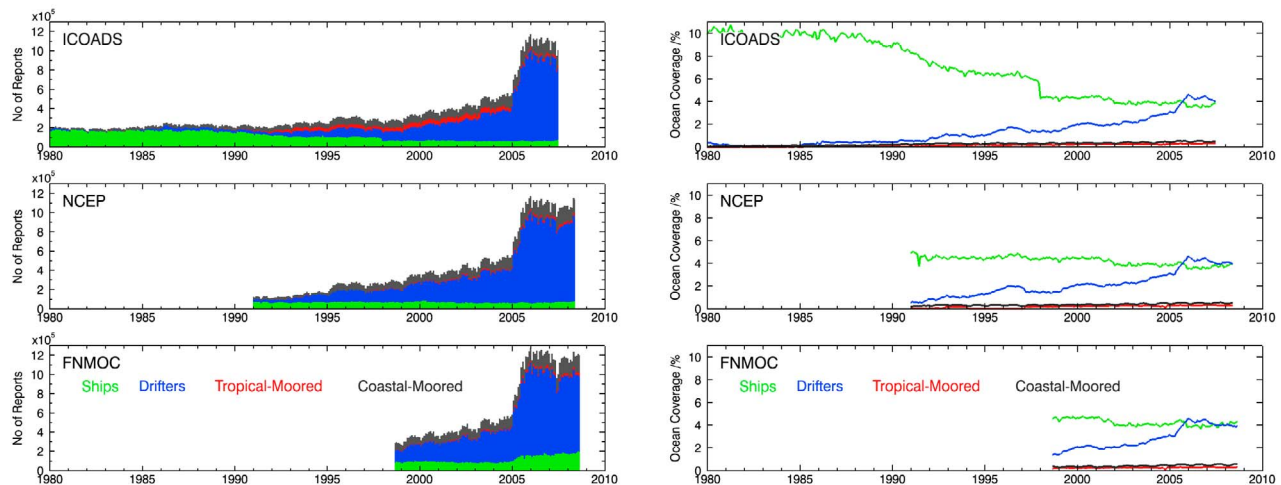
Figure 3. (left) Monthly number of SST reports corresponding to Figure 2. (right) Coverage of the ocean (60°S–60°N) by these reports (each point represents monthly median number of daily 1° × 1° boxes covered [cf. *Reynolds et al.*, 2002].

this writing, amount to ~100,000 to ~200,000 per month. At the same time, buoy reports have increased at a faster pace than the number of buoys and currently amount to ~1,000,000 reports per month. The most dramatic increase occurred in January 2005, thanks to a successful renegotiation of the joint tariff agreement between NOAA and the French operators of ARGOS service in 2004 (R. Lumpkin, personal communication, 2009). As a result, the satellite tracking system is now using five or six polar satellites, up from two [*Elipot and Lumpkin*, 2008].

[24] Coverage by in situ data (defined as a median number of daily 1° × 1° boxes covered, [*Reynolds et al.*, 2002]) is shown in Figure 3 (right). The coverage by ships was ~10% in the 1980s but has now declined to ~4%, while coverage by drifters has increased from a fraction of a percent to more than 4%. The low-volume reporting ships provide a comparable coverage with high-volume reporting drifters because ships travel faster than drifters. Coverage by moored buoys (both tropical and coastal) is an order of magnitude smaller compared to ships and drifters, even in recent years, because of their small number and fixed locations. Despite low volume of data, moored buoys are critically important for climate and weather research because of their unique geolocation and fixed positions.

2.5. Reference SST, T_{ref}

[25] In this study, a first-guess SST field, T_{ref}, is employed for two applications. First, it allows one to cross-evaluate the in situ SSTs from different platforms and data sets using consistent metrics of anomalies with respect to a reference SST, $T_{in\ situ} - T_{ref}$. Also, the SST anomalies are used to identify and remove outliers in situ SSTs as a part of the NESDIS QC described in section 3.3.

[26] Here, Reynolds optimal interpolation global 0.25° daily analysis SST ("AVHRR only") available from 1982 onward was selected as T_{ref} [*Reynolds et al.*, 2007]. Being a blended product of AVHRR satellite retrievals and quality-controlled ICOADS in situ SSTs, Reynolds SST is thus not fully independent from in situ data. As a result, the mean bias and standard deviation (SD) of $T_{in\ situ} - T_{ref}$ are not

representative of accuracy and precision of in situ data and are only used in this study for a relative evaluation of different data sets and platforms. Gridded 0.25° resolution Reynolds SST was bilinearly interpolated in space to each in situ observation. No interpolation in time was attempted as it would require a diurnal-cycle-resolved T_{ref}, which is currently unavailable.

[27] Another long-term reference SST is available from the 1.0° daily Pathfinder v5.0 data set from 1981 to 2008 (cloud screened, day/night average, quality level ≥7, accessible at ftp://data.nodc.noaa.gov/pub/data.nodc/pathfinder/). For brevity, all analyses below were done with respect to Reynolds SST except for section 5 in which Pathfinder SST was additionally used for cross-check.

3. Quality Control

[28] This section describes the Quality Flags available in ICOADS and Quality Indicators available in FNMOC data sets as well as the QC procedure currently employed at NESDIS with NCEP GTS data.

3.1. QF in ICOADS Data

[29] ICOADS release 2.4 data were downloaded in the International Maritime Meteorological Archive (IMMA) format without using any filters to exclude data. A comprehensive duplicate check was done in the DM archive prior to January 1998, mainly aimed at detecting ship reports that have been recorded in ICOADS twice (initially from GTS and later from digitized logbooks [*Slutz et al.*, 1985; http://icoads.noaa.gov/e-doc/other/dupelim_1980]). A scaled-down, duplicate elimination process was implemented in ICOADS from January 1998 to December 2004, when all ship reports came from GTS (see http://icoads.noaa.gov/e-doc/other/dupelim_1998). From January 2005 onward, only "exact match" duplicate elimination has been done (S. D. Woodruff, personal communication, 2009). A slot for a track check (using maximum moving speed) was reserved in the ICOADS data, but this flag is not set, partly because some data sources used

Table 2a. Standard Deviations of $T_{in\ situ} - T_{Reynolds}$ Labeled by ICOADS Trimming Flags for the Year 2000[a]

SD/K (%)	<2.8 Sigma	2.8–3.5 Sigma	3.5–4.5 Sigma	>4.5 Sigma	Trimming Inapplicable
Ships	1.22 (92.48)	3.03 (3.29)	4.18 (1.99)	10.17 (1.96)	5.40 (0.28)
Drifters	0.43 (95.39)	1.36 (1.11)	1.95 (0.66)	7.63 (0.65)	1.04 (2.19)
Tropical-moored	0.39 (98.48)	0.44 (1.14)	0.38 (0.30)	5.10 (0.08)	(0.00)
Coastal-moored	0.59 (98.80)	1.74 (0.60)	2.50 (0.19)	9.65 (0.18)	1.23 (0.23)

[a]The percentage of the corresponding reports is shown in brackets. Statistics are stratified by ships, drifters, tropical-moored buoys, and coastal-moored buoys.

in the DM archive have already applied track check beforehand (S. D. Woodruff, personal communication, 2009).

[30] Information from two major QF groups is available in the IMMA data: the trimming flags and the National Climatic Data Center (NCDC) QFs [*Slutz et al.*, 1985, Supp. J; *Worley et al.*, 2005; http://icoads.noaa.gov/e-doc/stat_trim]. Several other flags inherited from the original sources are also available, including the Canadian Marine Environmental Data Service QF (http://icoads.noaa.gov/e-doc/imma/imma_short.pdf).

[31] A trimming flag is calculated as an indicator of the normalized deviation of in situ SST from the background mean. The background mean is calculated as a running median within monthly $2° \times 2°$ boxes, and the deviation is normalized against the running lower- and upper-median deviations ("sigmas"). The flag can indicate up to 16 different combinations of quality conditions, including whether the SST measurement is falling within the ±2.8, ±3.5 and ±4.5 sigma limits, or whether the trimming is inapplicable because of such reasons as "limits missing" or "landlocked box" (http://icoads.noaa.gov/e-doc/stat_trim). As an example, Table 2a shows standard deviations of $T_{in\ situ} - T_{Reynolds}$ in year 2000 for the reports stratified into five different categories using the trimming flags provided in the ICOADS data. Apparently, the degradation of SST accuracy is well captured by the trimming flags. However, the fidelity of this approach critically depends on the number of observations available within a monthly $2° \times 2°$ box used to calculate the trimming limits. Moreover, some SST variations may be smoothed in space and time, thus suppressing the climate signal [*Wolter*, 1997; *Smith and Reynolds*, 2003]. A new set of adaptive QFs is currently being tested and might be available in the future releases of ICOADS data (http://icoads.noaa.gov/aqc.html). It will be digitized to 0.5 sigma steps and will be mainly aimed at avoiding filtering out large climate signals such as El Niño.

[32] The NCDC composite QF is an indicator of SST values outside long-term climatological limits [*Slutz et al.*, 1985, Supp. J]. It classifies reports into three categories: "correct" (within 4.8 sigma of the mean), "suspect" (4.8 to 5.8 sigma), and "erroneous" (>5.8 sigma), where mean and sigma are calculated as $2° \times 2°$ long-term monthly values.

Although the definitions of the mean and sigma in the NCDC QF are different from those used in the trimming flag, both flags are indicative of outliers in the data. Table 2b shows that the NCDC suspect and erroneous flags together identify less than 1% of data, and the corresponding SSTs are indeed anomalous. Note that the NCDC QF was considered in the duplicate elimination of ICOADS [*Slutz et al.*, 1985, Supp. J].

3.2. QI in FNMOC Data

[33] FNMOC provides a continuous QI ranging from 0 (best quality) to 1 (worst quality). As of this writing, no publicly accessible documentation on the QC procedure and definitions adopted in FNMOC data set was immediately available to the authors.

[34] Figure 4 shows two examples (for years 2000 and 2007) of FNMOC data characterization as a function of QI. Apparently, the QC strategy has changed between these 2 years. According to US GODAE (http://usgodae.fnmoc.navy.mil), there was a version upgrade in April 2004. However, additional analyses (not shown) indicate that there was another change (likely in 2006) before the new statistical patterns shown in Figure 4 (right) had finally established. In this third phase, the QI apparently has become more effective in discriminating measurements with high quality (small QI), whose fraction has significantly increased. At the same time, the bias and SD in the $T_{in\ situ} - T_{Reynolds}$ are now more closely associated with QI, and the tropical-moored buoys and drifters are apt to agree with Reynolds SST and with each other, followed by the coastal-moored buoys and ships.

3.3. NESDIS QC of NCEP GTS Data

[35] The current NESDIS QC employed with NCEP GTS data builds upon the existing QC procedures developed for different applications, including the ocean/atmosphere data assimilation [e.g., *Lorenc and Hammon*, 1988; *Ingleby and Huddleston*, 2007], SST optimal interpolation and averaging [*Smith and Reynolds*, 2003; *Rayner et al.*, 2006; *Stark et al.*, 2007], and data archival [*Slutz et al.*, 1985; *Baker*, 1994; *Woodruff et al.*, 1998; *Worley et al.*, 2005]. Satellite Cal/Val is very demanding in terms of the quality of in situ data and

Table 2b. Standard Deviations of $T_{in\ situ} - T_{Reynolds}$ Labeled for the ICOADS NCDC Composite QC Flags for the Year 2000[a]

SD/K (%)	Correct (<4.8 Sigma)	Suspect (4.8–5.8 Sigma)	Erroneous (>5.8 Sigma)
Ships	1.56 (98.90)	7.46 (0.38)	12.89 (0.72)
Drifters	0.54 (99.79)	3.50 (0.05)	12.06 (0.16)
Tropical-moored	0.39 (99.93)	0.55 (0.01)	4.13 (0.05)
Coastal-moored	0.62 (99.87)	7.34 (0.03)	10.95 (0.10)

[a]The percentage of the corresponding reports is shown in brackets. Statistics are stratified by ships, drifters, tropical-moored buoys, and coastal-moored buoys.
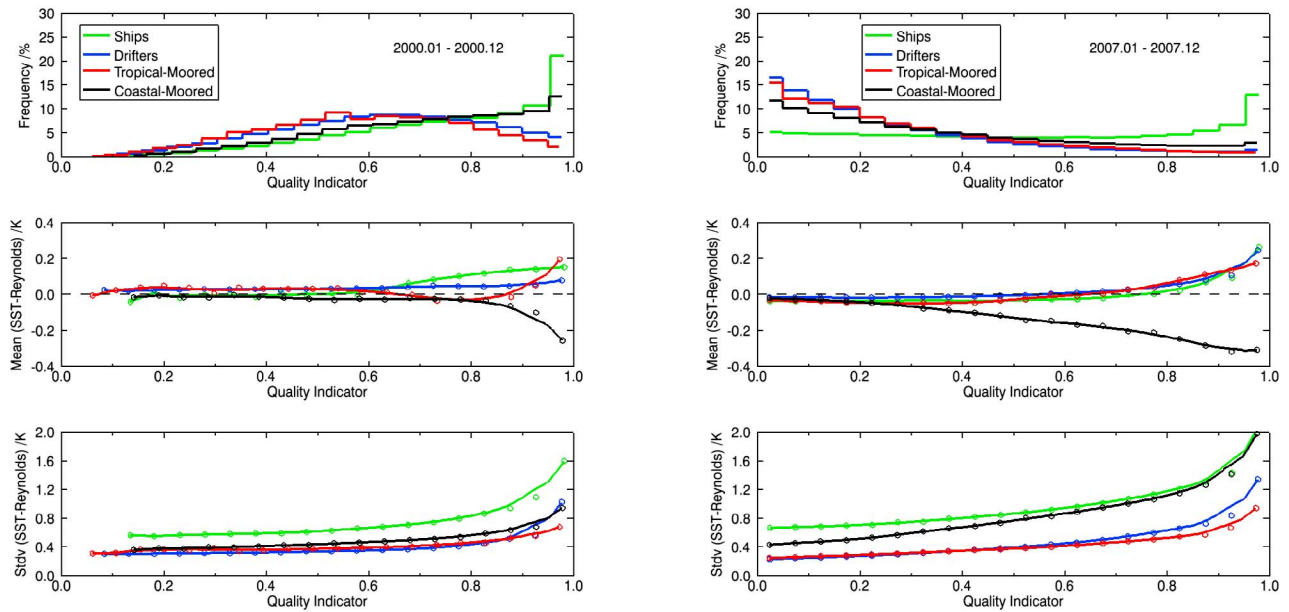
Figure 4. Characterization of FNMOC data as a function of Quality Indicator (QI). Note that the QI ranges from 0 (best quality) to 1 (worst quality). (top) Histograms of QI; (middle) mean bias in $T_{\text{in situ}} - T_{\text{Reynolds}}$; (bottom) SD of $T_{\text{in situ}} - T_{\text{Reynolds}}$ in years (left) 2000 and (right) 2007.

requires a flexible and scalable QC depending upon the specific task.

[36] QC checks are preceded by a geolocation check, which removes those reports whose latitude and longitude suggest that they are located over land or in the coastal areas (defined as <10 km from the coastline). Near-coast in situ data are removed because they cannot be accurately matched with the gridded reference SST because of the edge effects in the interpolation. Also, the corresponding in situ SSTs are highly variable in space and time, because of shallow waters and high dynamics, and should be avoided in satellite Cal/Val. The geolocation check eliminates ~8% of ship observations, ~8% of coastal-moored observations, ~1% of drifter observations, and ~0% of tropical-moored observations.

[37] Basic QCs follow, which include duplicate removal, track check, and outlier detection. Details are provided in the appendix. Figure 5 (left) shows the percentage of removed reports by each QC step. The duplicate check removes 2% to 4% of drifter and tropical-moored buoy reports. The track check works more effectively on moving (ships and drifters) rather than anchored (moorings) platforms and typically removes <0.1% of data. Outliers account for ~3% to 4% of data, except for tropical-moored buoys where the rate is ~1%. The total percentage of removed reports remains at an approximately constant level for ships and coastal moorings, while it varies significantly in time for drifters and tropical moorings because of a highly variable fraction of duplicates. Note that bursts in time series are artifacts due to a very small overall number of reports in some months.

[38] The SDs of $T_{\text{in situ}} - T_{\text{Reynolds}}$ in the quality-controlled reports are shown in Figure 5 (right). For reference, the SDs derived from all reports before QC are also superimposed. Typically, non-quality-controlled reports show abnormally high SDs, emphasizing the critical importance of QC. The effect is particularly large for drifters and tropical

moorings, which are the major source of in situ data used in satellite Cal/Val.

4. Comparisons Among Different Data Sets and Types

4.1. Applying NESDIS QC to All Data Sets

[39] For consistent analyses in the remainder of this paper, NESDIS QC described in section 3.3 was applied to all three data sets. For ICOADS and FNMOC, their own QC information was ignored and raw data were used as input to the NESDIS QC.

[40] Figure 6 shows time series of the percentage of eliminated reports in the ICOADS and FNMOC data, broken up by individual QC tests. The percentage of removed reports is fairly consistent across the three data sets for ships and moored buoys, but it significantly differs for drifters. Recall that duplicate records have been removed from the DM ICOADS data, which are largely duplicate free before December 2004. After ICOADS switched from DM to RT in January 2005, the fraction of duplicates becomes consistent with that in NCEP data. Apparently, the FNMOC processing does not remove duplicate records. Interestingly, beginning in ~2002, tropical moorings reported in FNMOC contain many more duplicates than the corresponding NCEP and ICOADS data.

[41] Data after the NESDIS QC are further analyzed in the remainder of this paper.

4.2. Statistics of $T_{\text{in situ}} - T_{\text{Reynolds}}$

[42] Time series of monthly mean biases and SDs with respect to Reynolds SST are plotted in Figure 7. Ship SSTs are biased warm by +0.1 to +0.2 K, likely because of the use of engine intake and thermometers on the VOS merchant ships. According to *Kent et al.* [1993] and *Kent and Taylor* [2006], warm biases are expected and may reach ~+0.35 K.
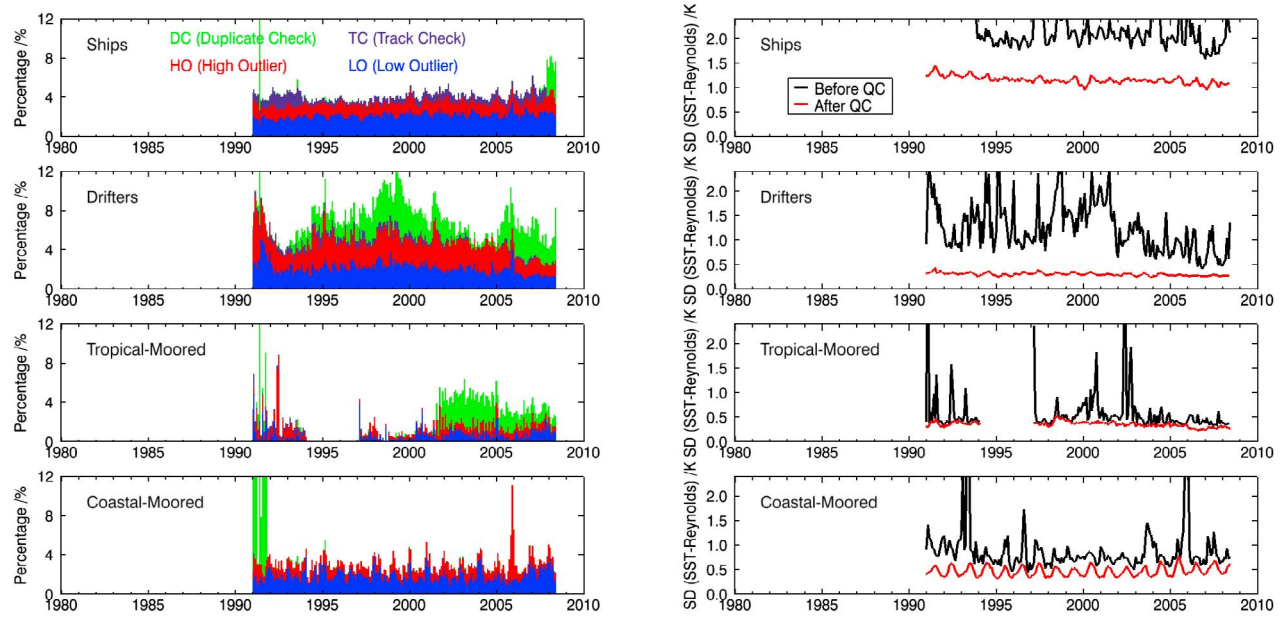
Figure 5. (left) Percent of reports identified by different NESDIS QC checks in the NCEP data (for explanation of data gap in tropical moorings from 1994 to 1997, see caption to Figure 1). (right) SD of $T_{\text{in situ}} - T_{\text{Reynolds}}$ comparison before QC versus after QC.

The corresponding SDs from 1.0 to 1.4 K likely result from different types of sensors employed on the ships and taking observations at a variety of depths ranging from the surface to around 25 m [*Kent et al.*, 1993; *Kent et al.*, 2007]. Also, the human recorders and data entry technicians tend to digitize SST to half-to-unit degrees and may introduce other large errors in the SST and ship position data during manual processing.

[43] Before approximately 1987, drifters were very few and their biases and SDs were unstable. Beginning in 1988, drifter bias has stabilized and now ranges from 0 to +0.05 K with corresponding SD ~0.3 to 0.4 K. Tropical-moored buoys have SD close to that of drifters and a mean bias from −0.05 to +0.10 K. Coastal-moored buoys have a comparable mean bias from −0.10 to +0.05 K but show strong seasonal variations in SD between 0.4 and 0.7 K. During the boreal winter, the statistics for coastal-moored buoys are close to those for drifters and tropical moorings, but they deteriorate significantly during the boreal summer. This seasonality will be discussed later in section 4.3.4 and section 5.2.

[44] All time series were noisy in early years because of scarcer statistics and have gradually improved. Ship SDs decreased in time since the early 1990s, likely because of the improvements in engine intake sensors [*Kent and Taylor*, 2006]. Similar trends are also captured by the long-term statistics with respect to Pathfinder SST (not shown). Mean biases and SDs of drifters and tropical-moored buoys indicate improved consistency with Reynolds SST after 2005, when more in situ data started being archived and began increasingly affecting the Reynolds SST.

4.3. Differences Among the NCEP, ICOADS, and FNMOC Data Sets

[45] To investigate differences among the three data sets, direct report-to-report comparisons have been performed.

For any two of the three data sets, the comparison approach is to first match the identical reports included in both data sets, and then to explore characteristics of the common and different portions. The matching conditions are the same as those used in the duplicate removal.

[46] Reports in different data sets could originate from the same platform, but they may have been received and/or processed differently. Differences in location (latitude and longitude) and time are expected to be within the digitizing precision, as no processing is intended to modify the original space-time information. In our matching process, ~0.1% of the reports were found to mismatch in SST by 0.1 to 0.3 K, although those records exactly match in space, time, and platform ID.

[47] For brevity, only results of NCEP versus ICOADS and FNMOC versus ICOADS are shown in Figure 8, from which differences between the NCEP and FNMOC can be indirectly inferred. Also, mean biases and SDs of the "intersection" and "complement" portions have been calculated for every pair of data sets (not shown). The major observations are summarized below.

4.3.1. Ships

[48] Most reports in different data sets originate from the same sources. Before January 1998, ICOADS DM archives included significantly more ship reports than NCEP archives. These additional reports are mostly found in the western tropical Pacific and belong to ships with IDs from 1339 to 7859. These ships employed hard copy reporting and did not report electronically via Argos and GTS [*Reynolds et al.*, 2002]. These extra reports do not show different quality in terms of mean bias and SD with respect to Reynolds SST.

[49] FNMOC consistently contains anywhere from 10,000 to 20,000 more ship reports per month than NCEP or ICOADS RT archives. These extra reports come from additional ships with five numeric digit call signs and show a
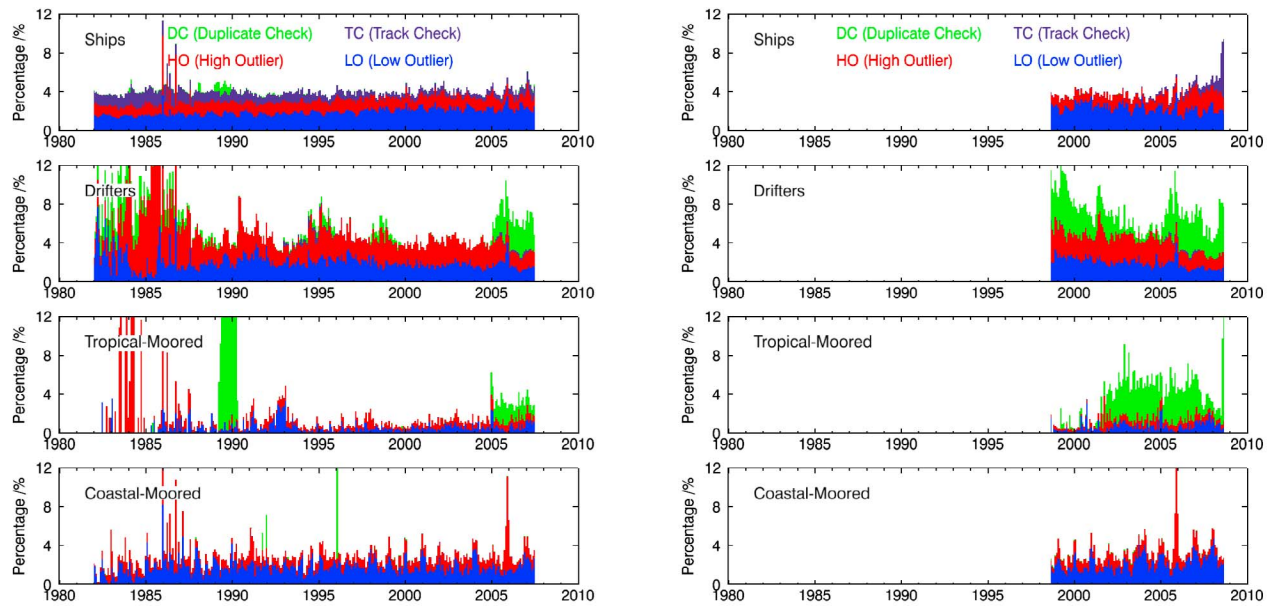
Figure 6. Percent of reports identified by different NESDIS QC checks in (left) ICOADS and (right) FNMOC.

stronger seasonal cycle in the mean bias but smaller SD. (Note that many extra FNMOC ships are found near coasts, and a considerable fraction of these extra reports was removed by geolocation check.)

[50] From October 1999 to April 2000, NCEP and FNMOC contain some 10,000 more ship reports than ICOADS. Geographical locations of these additional reports, along with low SD of ~0.6 K and cold bias of ~−0.1 K, are typical of moored buoys. They are frequently reported from some fixed positions in the coastal areas near the North America and in the Pacific. These reports are believed to be coastal mooring reports, which might have been erroneously labeled with a ship code. They were removed by the ICOADS duplicate elimination (cf. Section 2.1 of http://icoads.noaa. gov/e-doc/other/dupelim_1998).

4.3.2. Drifters

[51] Drifters appear very consistent across different data sets (in particular, FNMOC and NCEP data match very well before January 2007). ICOADS shows a somewhat smaller sample before January 2005 (because of stringent QC in RT processing), but the three data sets become identical beginning in 2005, when ICOADS switched to RT processing.

4.3.3. Tropical-Moored Buoys

[52] The major observation is that before December 2004, the frequency of reporting is much higher in ICOADS DM than in NCEP and FNMOC. This is because the ICOADS captures hourly reports while NCEP reports are sampled every 8 h or so.

4.3.4. Coastal-Moored Buoys

[53] Before 1995, ICOADS had twice as many records as NCEP because of much more frequent reporting. Since
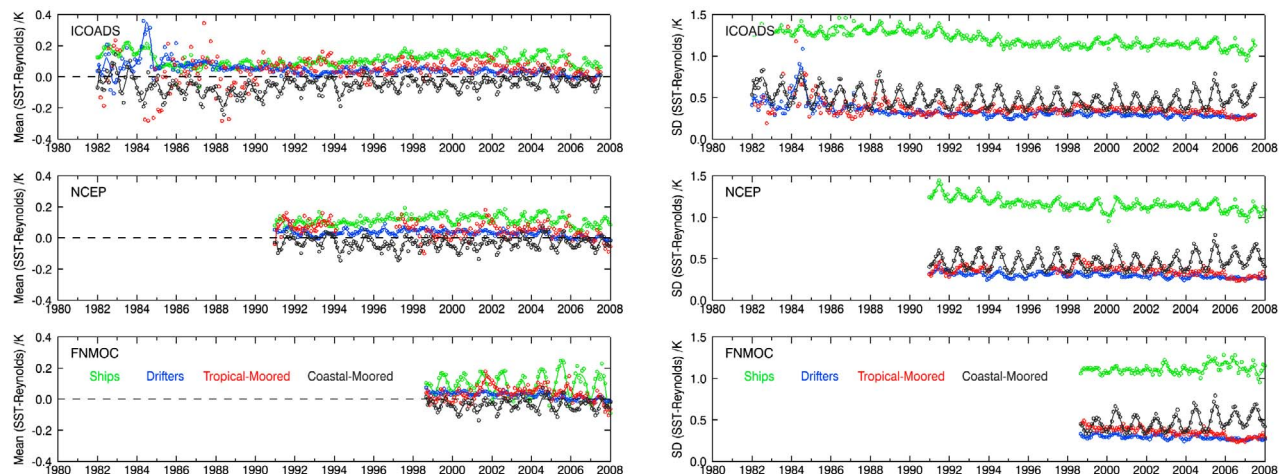

Figure 7. Time series of (left) monthly mean biases and (right) SD of $T_{\text{in situ}} - T_{\text{Reynolds}}$. Only data that pass NESDIS QC are used.
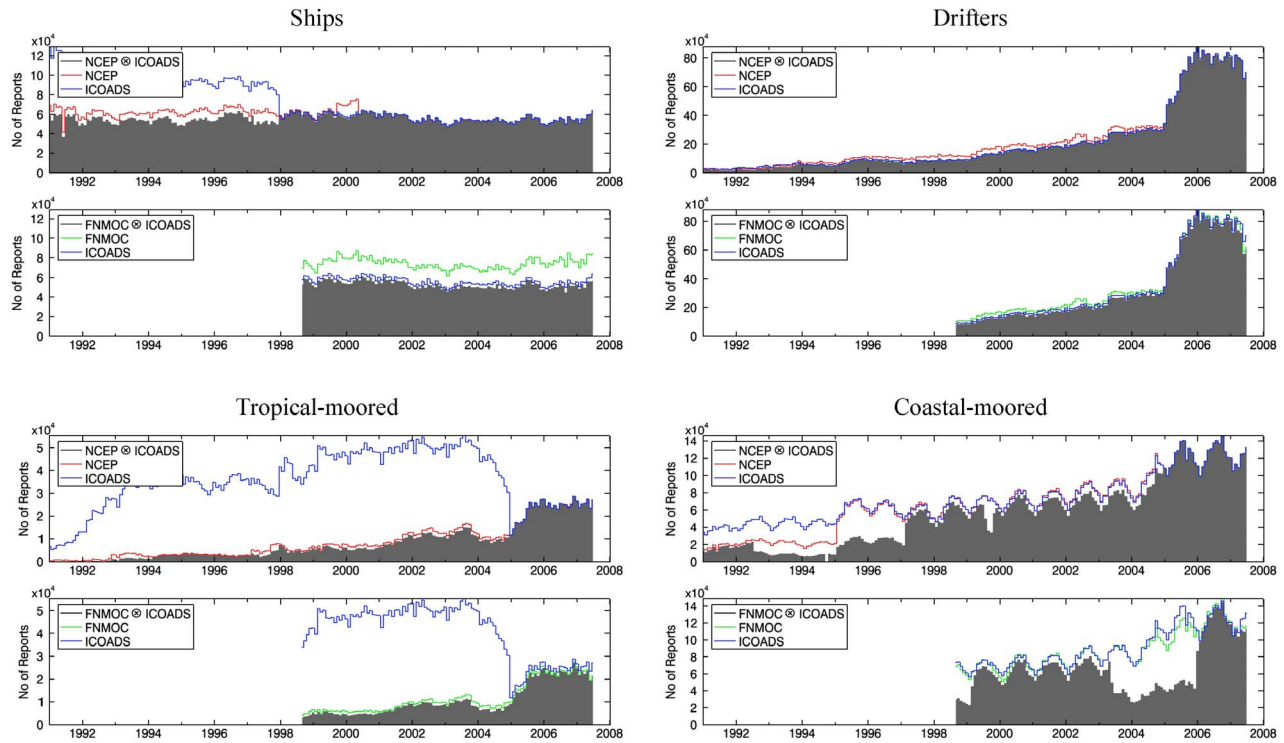
Figure 8. Comparisons of (top) NCEP versus ICOADS and (bottom) FNMOC versus ICOADS for different in situ data types. Note that the circled cross denotes intersection. For example, the circled cross between NCEP and ICOADS stands for a subset included in both NCEP and ICOADS.

January 1995, however, reporting density in NCEP has matched that of ICOADS.

[54] From June 1992 to February 1997, the common portion between NCEP and ICOADS decreased because a considerable number of reports did not match in time. Our additional analyses have shown that the time differences were very small and close to the threshold used here (0.1 h), and the "false mismatches" resulted from trivial differences in the ICOADS and NCEP data processing. The same problem also occurs in merging FNMOC and ICOADS between 2003 and 2006.

[55] Unlike the other three platforms, seasonal variations of the number of reports for the coastal-moored buoys are significant. The major reason is that some moorings are removed from the water in September–October for the freeze-up and redeployed in the spring (R. Crout, personal communication, 2009). A large part of these removed moorings are from the Great Lakes region and some other internal North American waters (with buoy IDs 45***). The seasonal operating period varies for different locations and years. This also partially explains the seasonal oscillations of the statistics of coastal moorings in Figure 7, i.e., measurements in inland waters usually have worse accuracies. It has been tested (not shown) that the removing of those "45***" buoys does help reduce the amplitude of seasonal cycles in Figure 7.

5. Characterization of In Situ SSTs for Satellite Cal/Val

[56] Representative coverage by in situ measurements in space and time, as well as small and uniform systematic and random errors, are critically important for satellite Cal/Val [*Emery et al.*, 2001]. Three subsections below analyze coverage, biases, and RMSEs in in situ SST observations, using 5 years of ICOADS data from January 2000 to December 2004. For these analyses, all data have been uniformly quality controlled using NESDIS QC described in section 3.3 and the Appendix.

[57] Analyses are performed with respect to two different T_{ref}: Reynolds and Pathfinder V5 1° product. Only best-quality Pathfinder day-night average data (with quality flag 7) were used [cf. *Reynolds et al.*, 2007]. The two T_{ref} values are used here to more thoroughly evaluate in situ SSTs. In addition, these results provide an independent validation of Pathfinder SST against in situ data. Recall that Reynolds product blends Pathfinder and in situ data by anchoring satellite SST to in situ SST, using a region- and season-specific bias correction. On the other hand, in Pathfinder, in situ SSTs are only used to calibrate the "global" satellite regression algorithms. This procedure centers satellite SST on in situ SST globally, but it fully preserves spatial contrasts in satellite retrievals [*Kilpatrick et al.*, 2001]. A moving-window bias correction is employed in Reynolds processing because satellite retrievals have variable biases. For instance, satellite infrared retrievals are made only under clear-sky conditions. As a result, Pathfinder SST is only available in a clear-sky domain and biased toward clear skies. At the same time, it may be subject to residual/ambient cloud contamination, uncorrected/under-corrected effects of aerosols, water vapor, and skin (satellite) minus bulk (in situ) differences [*Casey and Cornillon*, 1999].
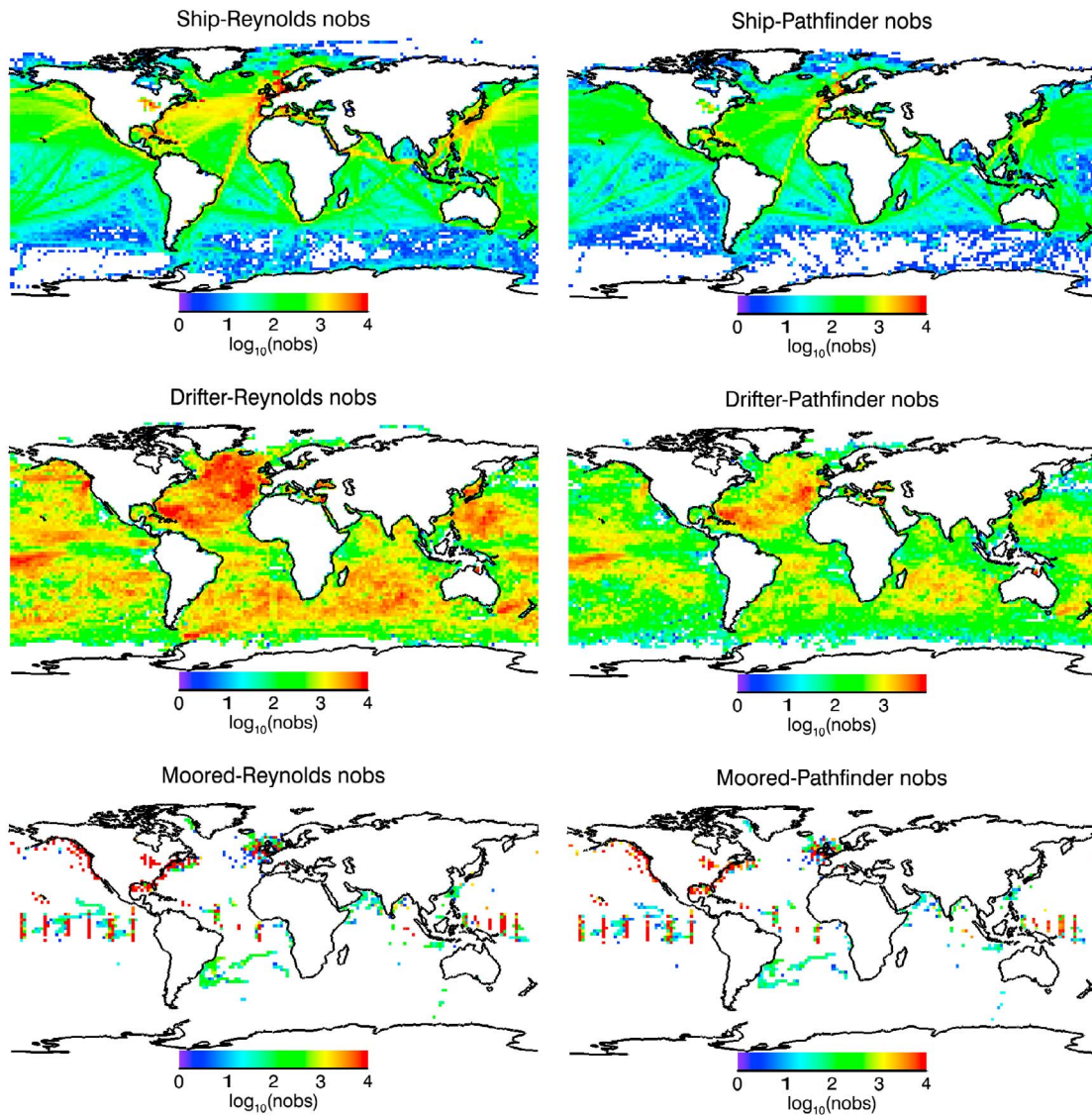
Figure 9. Global maps of in situ number of observation (NOBs) with respect to (left) Reynolds daily and (right) Pathfinder SST for a 5 year period of ICOADS data (January 2000–December 2004) in 2° × 2° boxes. (Note that the color bar is in log scale and the tropical- and coastal-moored buoys have been combined in one panel as they are well-separated geographically.)

5.1. Coverage by In Situ Data

[58] Figure 9 shows global densities of matchups with Reynolds and Pathfinder SSTs from the 5 years of ICOADS data aggregated in 2° × 2° boxes.

[59] The number of matchups with regular-gridded Reynolds data is representative of the true density of in situ data, whereas matchups with irregularly spaced (due to cloud screening) Pathfinder data are fewer than with Reynolds and only cover ~35% of in situ data.

[60] Of all platforms, drifters provide the densest and most complete global coverage. And yet their sampling is sparse and nonuniform in many areas. Ships well sample the Northern Hemisphere (with increased data density around major ship routes) but are very sparse in the tropics and especially in the Southern Hemisphere.

[61] Moored buoys are only available in some limited areas in the tropics and along coasts. Note that some points in the moored panels of Figure 9, e.g., in the central Pacific and the southern Atlantic, apparently belong to moving platforms. Additional analyses show that those are actually drifters that have been accidentally mislabeled as moored buoys. Their density of observations is an order of magnitude smaller than those from the moored platforms. Misclassifications likely exist for the drifters and ships, too, but they are more difficult to observe in Figure 9 because of higher density of data.

[62] To additionally analyze the global representativeness of in situ data, Figure 10 plots their zonal and SST densities. If in situ data cover the global ocean fully and uniformly, then their zonal and SST densities should closely match the
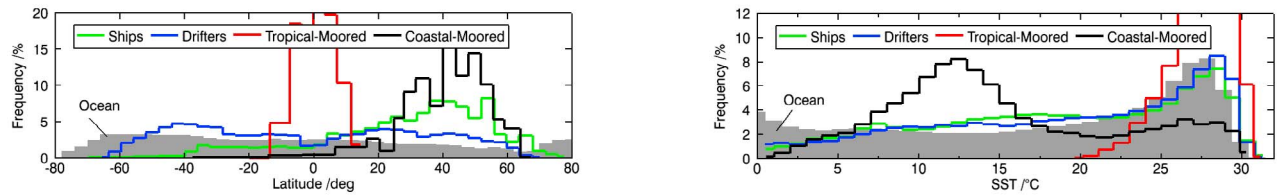
Figure 10. Histograms (integral normalized to 100%) as a function of (left) latitude (4° bin) and (right) SST (1°C bin) for different types of in situ data for 5 years of ICOADS data (January 2000–December 2004). Grey histograms show corresponding histogram of all ocean pixels. For best representation, different in situ data should closely match the ocean histograms.

corresponding all-ocean histogram (also shown in Figure 10 in grey). Drifters cover the full SST domain most uniformly except for the high latitudes (outside ∼±65° with SST <5°C) that are not covered at all, whereas the extratropical areas (outside ∼±10° latitude, corresponding to SSTs from ∼10°C

to 20°C) are slightly overrepresented. Some deficit of drifter data in the tropics is offset by the tropical moorings there, which by themselves do not represent global distribution of SST but can be combined with drifters and ships to improve the global representation of the matchup data sets. Ship data
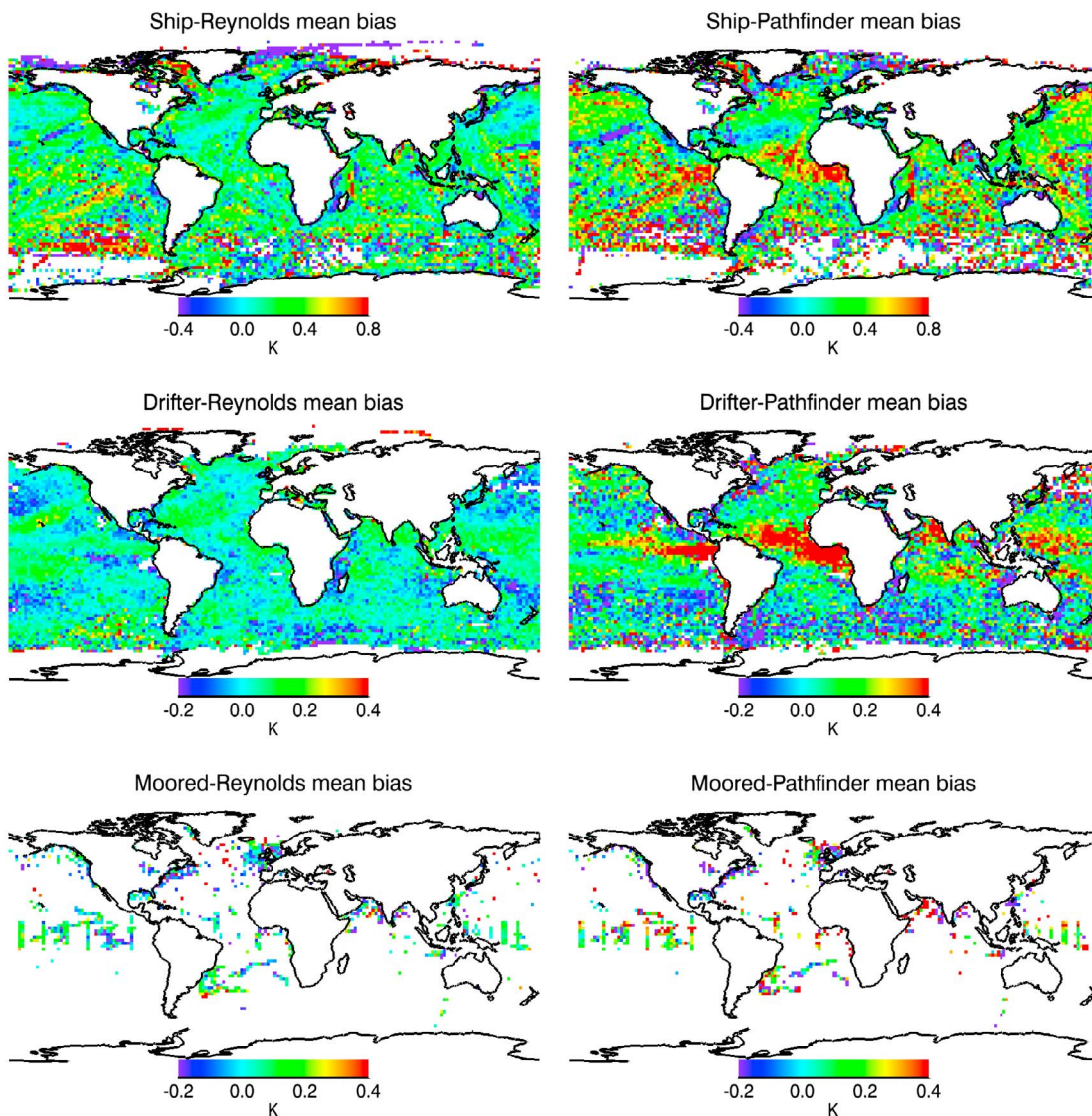

Figure 11. Same as in Figure 9 but for mean biases of (left) $T_{\text{in situ}} - T_{\text{Reynolds}}$ and (right) $T_{\text{in situ}} - T_{\text{Pathfinder}}$.
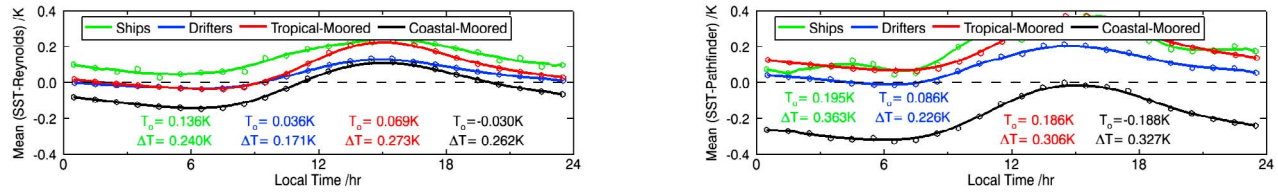
Figure 12. Mean biases in (left) $T_{\text{in situ}} - T_{\text{Reynolds}}$ and (right) $T_{\text{in situ}} - T_{\text{Pathfinder}}$ as a function of local time (1 h bin) with the average level T_0 and the dynamic range ΔT shown too. Data are 5 years of ICOADS from January 2000 to December 2004.

sample the SST dynamic range quite well, but their zonal coverage is very nonuniform. In particular, the Southern Hemisphere is strongly underrepresented (there are no data below 40°S), whereas the Northern Hemisphere is heavily overrepresented. Also, the ship data extend farther to the north than the drifter data. Results in Figure 10 suggest that the customary Cal/Val against in situ data may be biased toward well-populated geographical areas and corresponding SST domains but may be suboptimal in underrepresented data domains. Sampling strategies (based, for example, on histogram equalization of in situ observations and all ocean pixels) should improve the global representation of the Cal/Val results. Figure 10 also shows that there are no in situ data in the high latitudes and, therefore, no Cal/Val of satellite SST is possible in these areas. Additional analyses (not shown here) suggest that in situ data relatively uniformly cover the full diurnal and seasonal cycles, except for high-latitude coastal moorings, which are mostly employed in summer months because of water freezing.

5.2. Distribution of Mean Biases in Space and Time

[63] Figure 11 shows mean biases in $T_{\text{in situ}} - T_{\text{Reynolds}}$ and $T_{\text{in situ}} - T_{\text{Pathfinder}}$ corresponding to Figure 9.

[64] Because of bias correction employed in Reynolds product, the biases in $T_{\text{in situ}} - T_{\text{Reynolds}}$ are small and uniform in space (especially for drifting and moored buoys, which are used in Reynolds product with largest weight [*Reynolds et al.*, 2007]). The ship SSTs are biased warm as expected [cf. *Kent et al.*, 1993; *Kent and Taylor*, 2006].

[65] Pathfinder T_{ref} patterns are generally similar to those of Reynolds. However, there are some important differences. Pronounced warm biases in $T_{\text{in situ}} - T_{\text{Pathfinder}}$ up to +0.5 K are observed over some large areas, in particular in the tropics and high latitudes. Those are due to cold biases in Pathfinder SST in the areas of high and persistent cloud and aerosol load [*Casey and Cornillon*, 1999; *Reynolds et al.*, 2007]. Also, mean biases in $T_{\text{in situ}} - T_{\text{Pathfinder}}$ are noisier

than $T_{\text{in situ}} - T_{\text{Reynolds}}$ because Reynolds SST was smoothed in space using the optimal interpolation technique, whereas no spatial interpolation or smoothing was applied to Pathfinder data.

[66] Figure 12 plots mean biases as a function of local time. In situ SSTs are subject to diurnal warming and cooling, whereas Reynolds and Pathfinder SSTs do not resolve the diurnal cycle. The amplitude of the diurnal cycle with respect to Reynolds SST is only $\Delta T \sim 0.2 \pm 0.03$ K for ships and drifters but $\Delta T \sim 0.27$ K for moorings. The diurnal cycle is larger in the tropics, likely because of the higher insolation and lower winds, and in the coastal areas because of shallow waters. The diurnal amplitudes are larger with respect to Pathfinder because of sampling clear-sky areas in satellite SST associated with higher solar insolation and lower wind conditions.

[67] Figure 13 shows the same mean biases but as a function of month-of-year. When plotted with respect to Reynolds, the biases are close to zero for all platforms except ships, which are biased warm by ∼+0.1 to +0.2 K [cf. *Kent et al.*, 1993; *Kent and Taylor*, 2006]. The amplitude of seasonal changes in the bias is quite small for all in situ data. The largest signal, ∼0.13 K, is observed for coastal moorings, possibly because of different uncertainties in the Reynolds SST itself in the coastal areas and over internal waters in different seasons.

[68] Nonzero biases and larger seasonal variations are observed with respect to Pathfinder SST. Warm biases in ship, drifter, and tropical mooring SSTs are due to cold biases in Pathfinder SST itself. Figure 14 (left) additionally illustrates cold biases in Pathfinder with respect to Reynolds SST. Large negative biases in Pathfinder SST are seen, for example, in the tropical Atlantic and Pacific (in the areas associated with the Intertropical Convergence Zone and Saharan dust outbreak) and in the northern Pacific (in the areas associated with low stratus clouds, which are extremely difficult to detect in AVHRR imagery). Coastal moorings, on
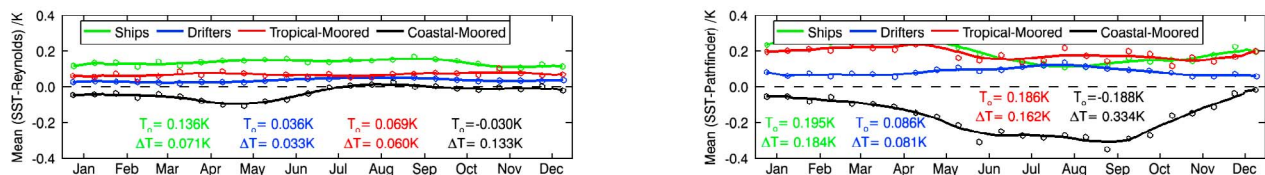

Figure 13. Mean biases in (left) $T_{\text{in situ}} - T_{\text{Reynolds}}$ and (right) $T_{\text{in situ}} - T_{\text{Pathfinder}}$ as a function of month-of-year (half-month bin) with the average level T_0 and the dynamic range ΔT shown too. Data are 5 years of ICOADS from January 2000 to December 2004.
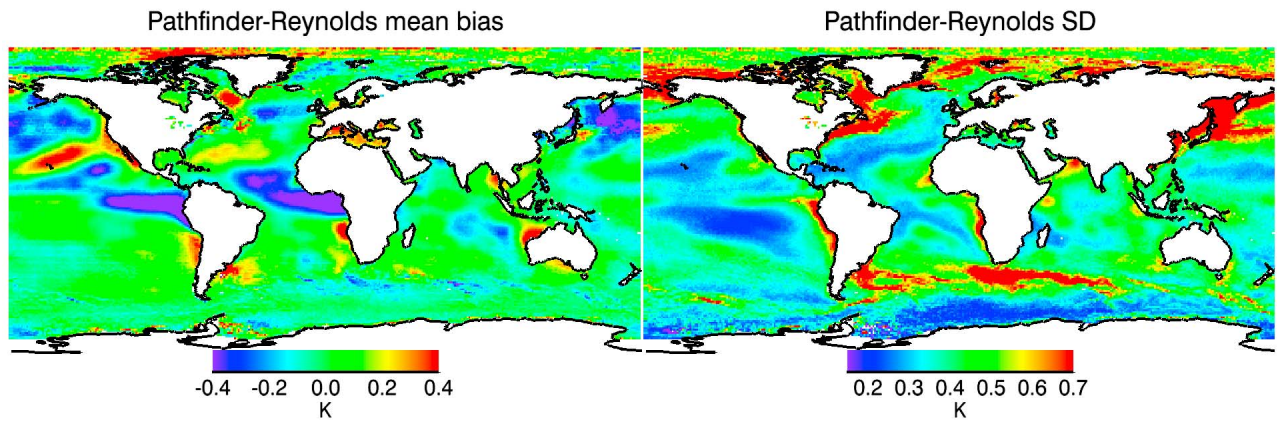
Pathfinder-Reynolds mean bias

Pathfinder-Reynolds SD

Figure 14. Global maps (2° × 2° resolution) of Pathfinder minus Reynolds SST (left) mean biases and (right) standard deviations for a 5 year period from January 2000 to December 2004.
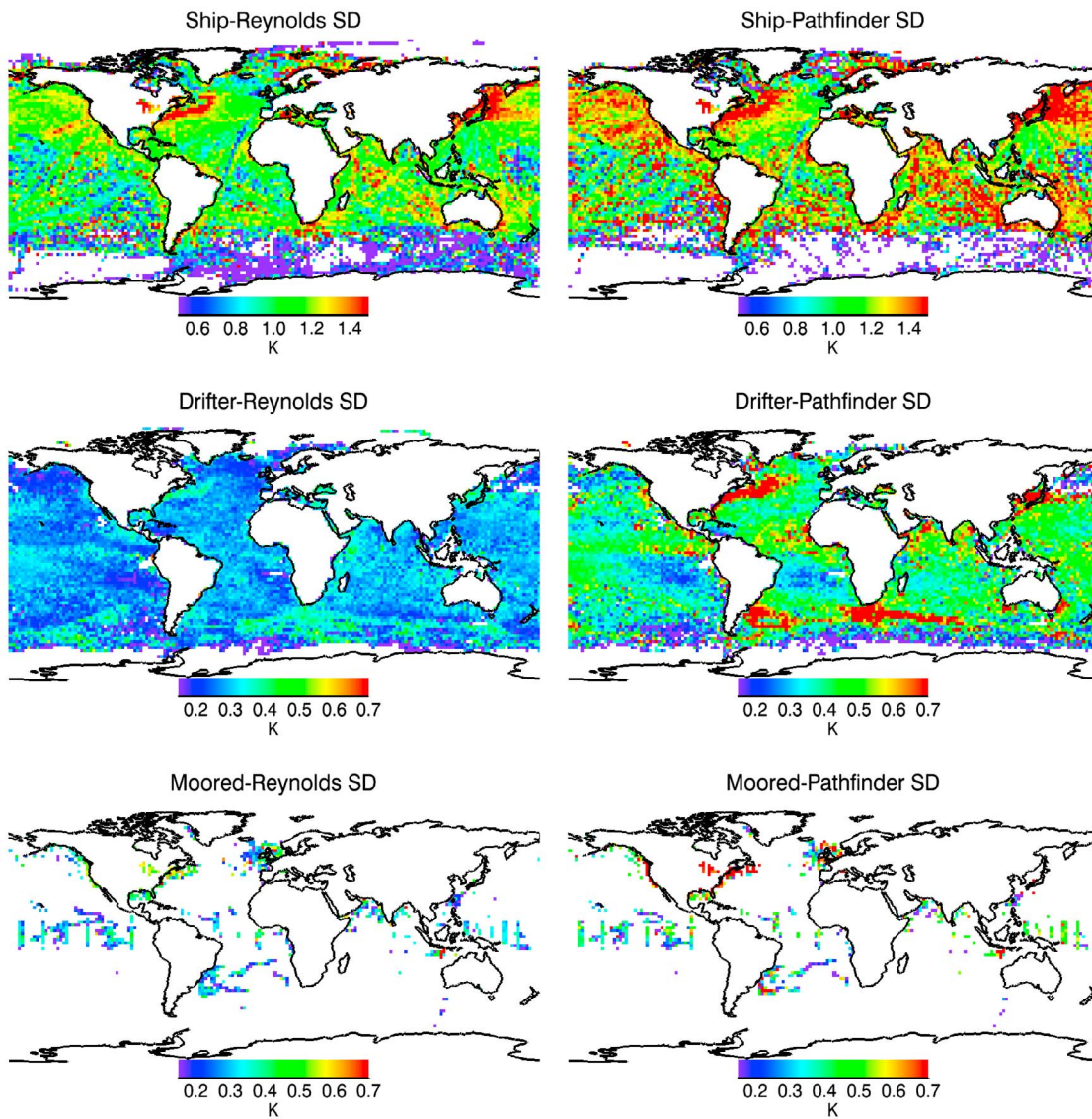
Ship-Reynolds SD

Ship-Pathfinder SD

Drifter-Reynolds SD

Drifter-Pathfinder SD

Moored-Reynolds SD

Moored-Pathfinder SD

Figure 15. Same as in Figure 9 but for standard deviations of (left) $T_{\text{in situ}} - T_{\text{Reynolds}}$ and (right) $T_{\text{in situ}} - T_{\text{Pathfinder}}$.
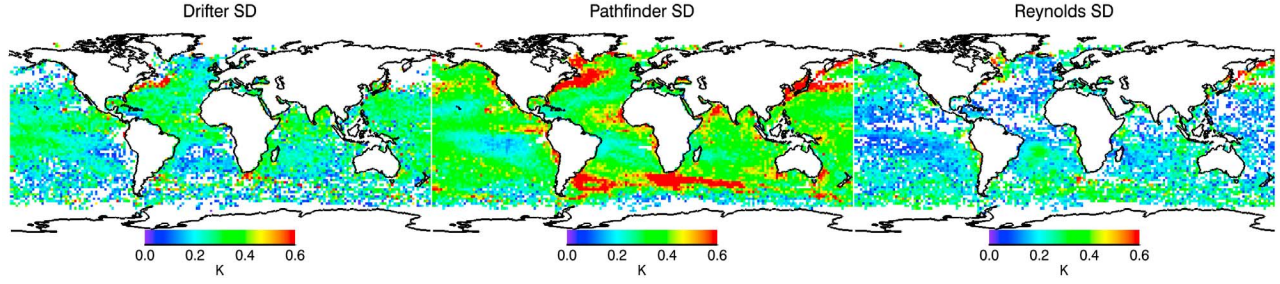
Figure 16. Global maps of root mean squared errors for (left) drifter, (center) Pathfinder, and (right) Reynolds SSTs derived using three-way error analysis [*O'Carroll et al.*, 2008].

the other hand, show a negative bias in $T_{in\ situ} - T_{Pathfinder}$, apparently because of a warm bias in Pathfinder SST (or cold biases in Reynolds SST) in the coastal areas and inland waters, which are apparently larger during the boreal summer.

5.3. Estimating RMSEs Using Three-Way Error Analysis

[69] For use in satellite Cal/Val, RMSE in in situ data should be small and uniform in space and time, or at least known and well characterized.

[70] Figure 15 shows SDs of $T_{in\ situ} - T_{Reynolds}$ (σ_{iR}) and $T_{in\ situ} - T_{Pathfinder}$ (σ_{iP}), and Figure 14 (right) additionally shows the SDs in $T_{Pathfinder} - T_{Reynolds}$ (σ_{PR}). Assuming that random errors in Reynolds (RMSE = σ_R), Pathfinder (σ_P), and in situ (σ_i; i = drifter, ship, tropical mooring, or coastal mooring) SSTs are uncorrelated, one can write three equations with three unknowns:

$$\sigma_{iR}^2 = \sigma_i^2 + \sigma_R^2; \quad \sigma_{iP}^2 = \sigma_i^2 + \sigma_P^2; \quad \sigma_{PR}^2 = \sigma_P^2 + \sigma_R^2 \quad (1)$$
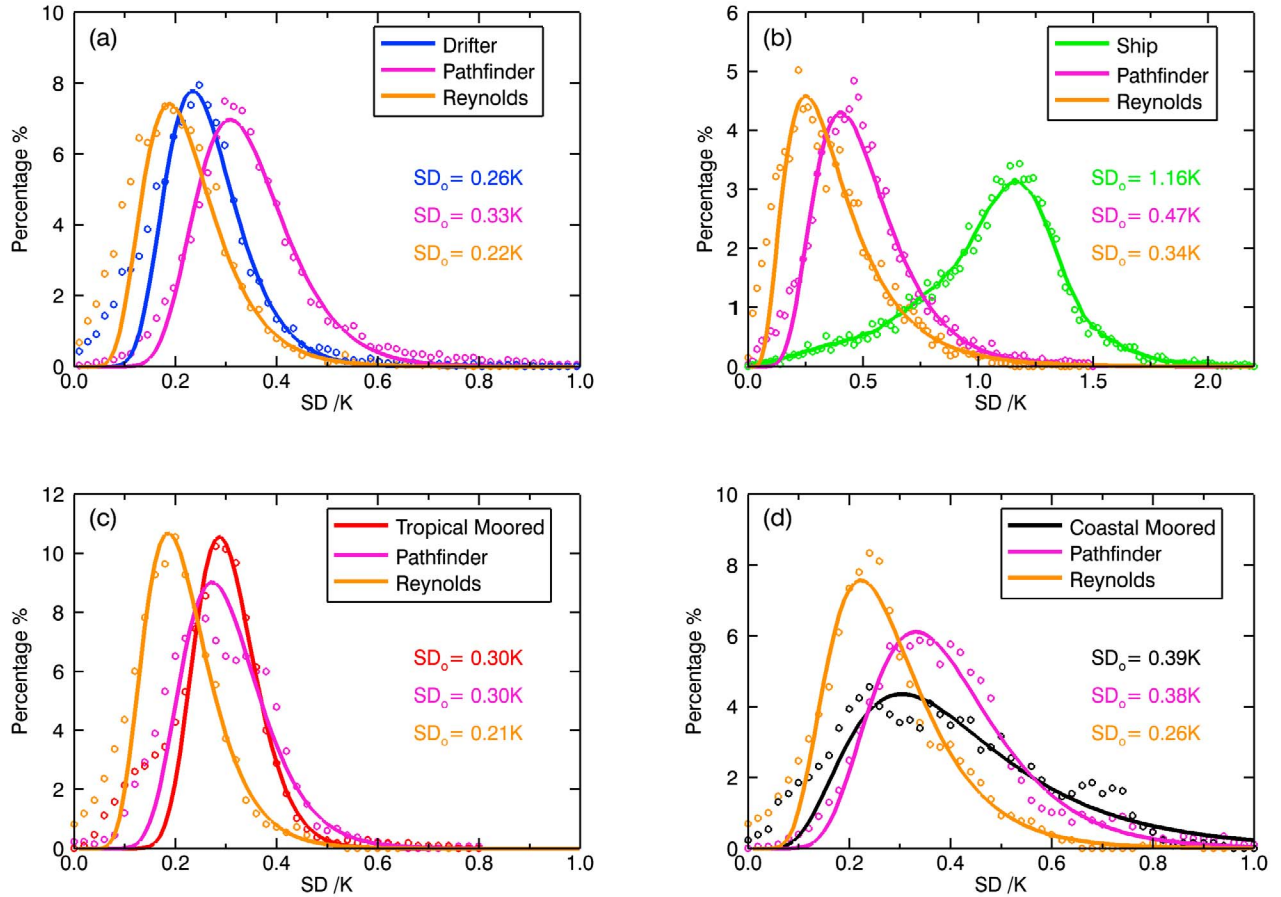

Figure 17. Histograms of RMSEs in Reynolds, Pathfinder, and different in situ data estimated using three-way error analysis [*O'Carroll et al.*, 2008] (for details, see section 5.3).

and solve them for σ_R, σ_P, and σ_i. *O'Carroll et al.* [2008] proposed this "three-way error analysis" method and applied it to data from two satellite sensors (Advanced Along-Track Scanning radiometer (AATSR) and Advanced Microwave Scanning Radiometer (AMSR-E)) and drifting buoys. They came up with the following estimates of global RMSEs in the three data sets: $\sigma_{AATSR} = 0.16$ K, $\sigma_{AMSR-E} = 0.42$ K, and $\sigma_i = 0.23$ K.

[71] Likewise, the three-way error analysis was applied here to estimate σ_R, σ_P, and σ_i. Note that the assumption that errors are uncorrelated is more valid for Pathfinder and in situ SSTs then for Reynolds SST, which is their blended product. As a result, the RMSEs estimated here may not be fully accurate, but some additional consistency checks suggest that they are realistic.

[72] Figures 14 and 15 show that σ_{iR}, σ_{iP} and σ_{PR} vary in space. Therefore, equations have been solved in each $2° \times 2°$ box, and the result is shown in Figure 16 in a form of three maps of σ_i, σ_P, and σ_R. In several boxes, either σ_R^2, σ_P^2, or σ_i^2 came out slightly negative, likely because of errors in the data or violation of the noncorrelation assumption; the respective boxes in Figure 16 are rendered as blank. All three σ values have a complex spatial structure, especially σ_P.

[73] Figure 17a additionally plots histograms of σ_R, σ_P, and σ_i from Figure 16. The value of σ_i varies from 0 to 0.6 K with a median of ~0.26 K. These estimates are in good agreement: the estimates of 0.23 K by *O'Carroll et al.* [2008] and 0.40 K by *Emery et al.* [2001]. For Pathfinder, median σ_P is ~0.33 K, but the histogram has a long tail extending out to 0.8 K and even beyond. Figure 16 confirms that there are large areas where Pathfinder SST has large RMSEs, and those are typically associated with persistent clouds and aerosols. For the blended Reynolds SST, the median $\sigma_R \sim 0.22$ K is smaller than both σ_P and σ_i. This is expected because a blended product should improve upon the RMSEs of the two inputs.

[74] Figures 17b, 17c, and 17d show results of corresponding three-way analyses for the other in situ data types. For tropical-moored buoys, the σ_i values are comparable with drifters (median $\sigma_i \sim 0.30$ K) whereas for coastal moorings, the errors are larger and more variable (median $\sigma_i \sim 0.39$ K). Ship SSTs are very noisy and their characteristics are very nonuniform, with a median σ_i of ~1.16 K. Note also that the median σ_R ranges from 0.21 to 0.34 K, and the median σ_P ranges from 0.30 to 0.47 K in different panels of Figure 17, likely because of the differences in the corresponding spatial samplings and violation of the noncorrelation assumption.

[75] Note that the numbers estimated here pertain to the $2° \times 2°$ average data. RMSEs for individual in situ measurements and Pathfinder retrievals, as well as for $0.25°$ Reynolds data, are likely larger, and separate analyses are needed to estimate them.

6. Conclusion

[76] This study was devoted to the analyses of in situ SST data and metadata, including different available data sets and types, their QC, representation for satellite Cal/Val, and error characteristics. There is a growing consensus in the SST community that such inventory is prerequisite to any

geophysical data analysis and improved Cal/Val of satellite SST products from the current and future sensors and is also prerequisite to reprocess historical data and generate improved climate data records.

[77] Cross-evaluation of NCEP GTS, FNMOC, and ICOADS data shows that most records in different data sets match most of the time. Compared to NCEP GTS, the FNMOC consistently includes more reports from some additional ships, and the ICOADS DM archive contains more complete records for ships and moored buoys. For the FNMOC data set to be used for satellite Cal/Val applications, periodic reprocessing data using the latest and consistent QC procedure and extending data record back in time to the beginning of satellite era (circa 1981) is needed, along with documenting its quality indicators. Being a long-term, most complete, well-organized, and well-documented data set, the ICOADS is well suited for satellite reprocessing efforts such as Pathfinder [*Kilpatrick et al.*, 2001]. The current efforts by the ICOADS team toward a more uniform and up-to-date DM archive throughout the satellite era will greatly facilitate the use of ICOADS data in satellite Cal/Val. For near-real time (NRT) SST Cal/Val work at NESDIS in the foreseeable future, NCEP GTS data will continue to be used, which requires an accurate and flexible QC.

[78] The quality of in situ SSTs is highly nonuniform, and QC is critically important before they can be used in satellite Cal/Val. QC information is available in ICOADS and FNMOC data sets. However, it is not always uniform and consistent in time (ICOADS, FNMOC) or not fully documented (FNMOC). Improvements to the QC system for in situ SSTs are underway at NESDIS. One of the priorities is achieving consistency between the QC procedures adopted in the remote sensing community and those employed in the meteorology and oceanography communities, e.g., in ICOADS [*Slutz et al.*, 1985; *Woodruff*, 2008] and the UK Met Office [*Lorenc and Hammon*, 1988; *Ingleby and Huddleston*, 2007].

[79] Our analyses confirm the previous observations by *Strong and McClain* [1984] and *Emery et al.* [2001] that drifting buoys offer the best global coverage and quality. Their biases measured with respect to Reynolds SST do not exceed several hundredths of a kelvin globally, and their RMSEs range from 0 to 0.6 K, with modal value $\sigma_{im} \sim 0.26$ K. Tropical-moored buoys cover a narrow domain within $\pm 20°$ latitude and have only slightly larger biases and RMSEs, with $\sigma_{im} \sim 0.30$ K. They are not representative of global SST, but can be used in concert with drifters for satellite Cal/Val. Coastal-moored buoys show a strong seasonality in number of observations, biases, and noise. Their use for satellite Cal/Val may only be recommended on a case-by-case basis. Ships cover the SST dynamic range well, but their geographical coverage is very nonuniform. Furthermore, they are biased ~+0.14 K warm with respect to Reynolds SST and show large RMSEs, with $\sigma_{im} \sim 1.16$ K. Although ship data are less accurate than drifters, they are the major source of in situ data in the pre-1990 period. Analyses are therefore needed to cross-evaluate the SST Cal/Val using ships and buoys when both data are available in sufficient quantities and to harmonize their use for historical reprocessing.

[80] Neither in situ data source is fully representative of the global distribution of SST and the coverage changes in time. Therefore, subsampling strategies should be employed

to form representative matchup data sets for the use in satellite Cal/Val. Also, neither in situ data type is fully accurate. Analyses toward estimating error budget in satellite and in situ SSTs from the data itself, similar to those proposed by *Emery et al.* [2001] and *O'Carroll et al.* [2008], should continue. An example of such self- and cross-consistency analyses was given in this study to estimate error budget in Pathfinder, Reynolds, and in situ data. It suggests that RMSEs in Pathfinder SST have a distinct spatial structure, with error ranging from ~0.1 to 1.0 K, and modal values from 0.30 to 0.47 K for $2° \times 2°$ averaged product. Our analyses also confirm the earlier observations by *Casey and Cornillon* [1999] and *Reynolds et al.* [2007] that Pathfinder SST has large areas of cold biases associated with residual cloud and aerosol contamination, particularly in the tropics. Warm biases in "Pathfinder minus Reynolds" are also observed in the coastal areas and inland waters, mostly during summertime. These biases and nonuniformities in RMSEs should be reconciled in the future SST climate data records (e.g., Pathfinder) and SST analyses (e.g., Reynolds).

Appendix A: NESDIS QC of In Situ SSTs

[81] The first step is duplicate removal. Reports that are close enough in space and time are considered duplicates. The conditions are latitude and longitude difference of <0.01° and time difference of <0.1 h. This step may also serve for data thinning [*Ingleby and Huddleston*, 2007]. For a group of "duplicates," the one with the best quality is kept. If the quality information is not available and if all the duplicates have SSTs within 0.1°C tolerance, then the first in the sequence is kept and the rest are eliminated; otherwise, all of them are eliminated.

[82] Track checking is next. It checks reports from the same platform for space-time consistency assuming a maximum moving speed. The whole track of reports from the same platform (ID) within 1 month is checked together. A least-required speed is calculated, assuming that the platform had traveled between the locations of any two reports through a direct linked path compared to a maximum possible speed for this type of platform, and anomalous reports are identified and excluded.

[83] Additional analyses have shown that track check is not very sensitive to the selection of maximum speed. Even a small error in a position will result in an extremely large speed. Therefore, maximum speed is liberally chosen as 120 km/h for ships or 80 km/h for drifters. A generous margin is allowed to minimize erroneous removal of good reports, which may result from insufficient accuracy of longitude, latitude, or time stamp. The moored buoys are supposed to be located in the same position at all times. However, reports with erroneous locations are still found. In this case, reports located far away from the majority of reports will be identified as erroneous. Note that reports with a group ID (several platforms that share the same ID) are not subject to track check.

[84] In the initial implementation of QC adopted in this study, outlier detection is done by comparing in situ SSTs to reference SSTs. Given Reynolds SST as the reference, measurements falling outside the "median ± 4 sigma" range are considered outliers. Sigma is calculated as a scaled Median Absolute Deviation (MAD), where the scale factor is selected as 1.4826 so that sigma will match SD for a perfect Gaussian distribution. The global median and sigma are calculated monthly from the $T_{\text{in situ}} - T_{\text{Reynolds}}$ anomaly separately for ships, drifters, and tropical- and coastal-moored buoys. The outlier test is anticipated to exclude the remaining erroneous reports with wrong SST, time, latitude, and/or longitude values due to operation, transmission, and processing errors and to exclude abnormal reports from dysfunctional or extremely noisy sensors.

[85] Note that this simple procedure, which is widely employed for QC in a postprocessing mode (see, for example, *Kent et al.* [1993]), has its own limitations. In particular, reference SST may not always be available in real-time application. Also, real-time QC of in situ data is supposed to be more robust and independent, so that no vicious circle is caused by mutual referencing or mutual utilization of in situ and reference SSTs. Most importantly, using one global monthly number for median and MAD for screening may result in overscreening in the dynamic areas (such as the Gulf Stream) or under-screening in the stable areas (such as the tropics). Also, using one set of screening criteria for in situ data collected during day and night and referencing them to Reynolds SST, which does not resolve diurnal cycle, likely results in suppressed diurnal cycle.

[86] More advanced and independent QC methods, such as Bayesian-based QC [*Lorenc and Hammon*, 1988], are being explored and will be employed in the future.

References

Baker, N. L. (1994), Quality control of meteorological observations at Fleet Numerical Meteorology and Oceanography Center, Nav. Res. Lab., Monterey, Calif., Dec.

Bourles, B., et al. (2008), The PIRATA program: History, accomplishments, and future directions, *Bull. Am. Meteorol. Soc.*, *89*, 1111–1125.

Casey, K., and P. Cornillon (1999), A comparison of satellite and in situ–based sea surface temperature climatologies, *J. Clim.*, *12*(6), 1848–1863.

Donlon, C., et al. (2007), The Global Ocean Data Assimilation Experiment High-resolution Sea Surface Temperature Pilot Project, *Bull. Am. Meteorol. Soc.*, *88*(8), 1197–1213.

Elipot, S., and R. Lumpkin (2008), Spectral description of oceanic near-surface variability, *Geophys. Res. Lett.*, *35*, L05606, doi:10.1029/2007GL032874.

Emery, W. J., D. J. Baldwin, P. Schlussel, and R. W. Reynolds (2001), Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements, *J. Geophys. Res.*, *106*(C2), 2387–2405.

Ingleby, B., and M. Huddleston (2007), Quality control of ocean temperature and salinity profiles – Historical and real-time data, *J. Mar. Syst.*, *65*, 158–175.

Kent, E. C., and P. G. Challenor (2006), Toward estimating climatic trends in SST. Part II: Random errors, *J. Atmos. Oceanic Technol.*, *23*(3), 476–486.

Kent, E. C., and P. K. Taylor (2006), Toward estimating climatic trends in SST. Part I: Methods of measurement, *J. Atmos. Oceanic Technol.*, *23*(3), 464–475.

Kent, E. C., P. K. Taylor, B. S. Truscott, and J. S. Hopkins (1993), The accuracy of voluntary observing ship's meteorological observations: Results of the VSOP-NA, *J. Atmos. Oceanic Technol.*, 10, 591–608.

Kent, E. C., P. G. Challenor, and P. K. Taylor (1999), A statistical determination of the random observational errors present in voluntary observing ships meteorological reports, *J. Atmos. Oceanic Technol.*, 16(7), 905–914.

Kent, E. C., D. I. Berry, S. D. Woodruff, and P. K. Taylor (2006), Voluntary observing ships: A vital observing system in decline, *CLIVAR Exchanges, No. 38,* 11(3), 20–21.

Kent, E. C., S. D. Woodruff, and D. I. Berry (2007), WMO publication no. 47 metadata and an assessment of voluntary observing ships observation heights in ICOADS, *J. Atmos. Oceanic Technol.*, 24(4), 214–234.

Kilpatrick, K. A., G. P. Podesta, and R. Evans (2001), Overview of the NOAA/NASA advanced very high resolution radiometer Pathfinder algorithm for sea surface temperature and associated matchup database, *J. Geophys. Res.*, 106(C5), 9179–9197.

Lorenc, A. C., and O. Hammon (1988), Objective quality control of observations using Bayesian methods. Theory, and a practical implementation, *Q. J. R. Meteorol. Soc.*, 114, 515–543.

McClain, E. P., W. G. Pichel, and C. C. Walton (1985), Comparative performance of AVHRR-based multichannel sea surface temperatures, *J. Geophys. Res.*, 90(C6), 11,587–11,601.

McPhaden, M., et al. (1998), The Tropical Ocean-Global Atmosphere observing system: A decade of progress, *J. Geophys. Res.*, 103(C7), 14,169–14,240.

McPhaden, M. J., et al. (2009), RAMA: The Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction, *Bull. Am. Meteorol. Soc.*, 90(4), 459–479.

O'Carroll, A. G., J. R. Eyre, and R. W. Saunders (2008), Three-way error analysis between AATSR, AMSR-E, and in situ sea surface temperature observations, *J. Atmos. Oceanic Technol.*, 25(7), 1197–1207.

Rayner, N. A., P. Brohan, D. E. Parker, C. K. Folland, J. J. Kennedy, M. Vanicek, T. J. Ansell, and S. F. B. Tett (2006), Improved analyses of changes and uncertainties in sea surface temperature measured in situ since the mid-nineteenth century: The HadSST2 data set, *J. Clim.*, 19, 446–469.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang (2002), An improved in situ and satellite SST analysis for climate, *J. Clim.*, 15, 1609–1625.

Reynolds, R. W., T. M. Smith, C. Liu, D. B. Chelton, K. S. Casey, and M. G. Schlax (2007), Daily high-resolution blended analyses for sea surface temperatures, *J. Clim.*, 20, 5473–5496.

Servain, J., et al. (1998), A Pilot Research Moored Array in the Tropical Atlantic (PIRATA), *Bull. Am. Meteorol. Soc.*, 79, 2019–2031.

Slutz, R. J., S. J. Lubker, J. D. Hiscox, S. D. Woodruff, R. L. Jenne, D. H. Joseph, P. M. Steurer, and J. D. Elms (1985), *Comprehensive Ocean-Atmosphere Data Set; Release 1*, http://icoads.noaa.gov/Release_1/coads.html, NOAA Environ. Res. Lab., Clim. Res. Prog., Boulder, Colo.

Smith, T. M., and R. W. Reynolds (2003), Extended reconstruction of global sea surface temperatures based on COADS data (1854–1997), *J. Clim.*, 16, 1495–1510.

Stark, J. D., C. J. Donlon, M. J. Martin, and M. E. McCulloch (2007), An operational, high resolution, real time, global sea surface temperature analysis system, paper presented at Oceans 2007: Marine Challenges: Coastline to Deep Sea, IEEE/Oceanic Eng. Soc., Aberdeen, Scotland, 18–21 June.

Strong, A. E., and E. P. McClain (1984), Improved ocean surface temperatures from space – Comparisons with drifting buoys, *Bull. Amer. Meteorol. Soc.*, 65(2), 138–142.

Walton, C. C., W. G. Pichel, J. F. Sapper, and D. A. May (1998), The development and operational application of nonlinear algorithms for the measurement of sea surface temperatures with the NOAA polar-orbiting environmental satellites, *J. Geophys. Res.*, 103, 27,999–28,012.

Wolter, K. (1997), Trimming problems and remedies in COADS, *J. Clim.*, 10, 1980–1997.

Woodruff, S. D. (2008), Common issues of quality control of surface marine data, paper presented at Data Management Coordination Group 3rd Session, JCOMM, Ostend, Belgium, 26–28 March.

Woodruff, S. D., H. F. Diaz, J. D. Elms, and S. J. Worley (1998), COADS release 2 data and metadata enhancements for improvements of marine surface flux fields, *Phys. Chem. Earth*, 23, 517–527.

Woodruff, S. D., H. F. Diaz, E. C. Kent, R. W. Reynolds, and S. J. Worley (2008), The evolving SST record from ICOADS, in *Climate Variability and Extremes during the Past 100 Years*, vol. 33, *Advances in Global Change Research*, edited by S. Brönnimann et al., pp. 65–83, Springer, New York.

World Meteorological Organization (2007), *WMO No. 386, Manual on the Global Telecommunication System*, World Meteorol. Organ., Geneva, Switzerland.

Worley, S. J., S. D. Woodruff, R. W. Reynolds, S. J. Lubker, and N. Lott (2005), ICOADS release 2.1 data and products, *Int. J. Climatol.*, 25(7), 823–842.

Worley, S. J., et al. (2009), The role of ICOADS in the sustained ocean observing system, community white paper presented at OceanObs'09, Intergov. Oceanogr. Comm. (UNESCO) and Eur. Space Agency, 21–25 Sept.

A. Ignatov and F. Xu, Center for Satellite Applications and Research, National Environmental Satellite, Data and Information Service, NOAA, Camp Springs, MD 20746, USA. (feng.xu@noaa.gov)